Overview of the TREC 2009 Web Track

Charles L. A. Clarke
University of Waterloo

Nick Craswell
Microsoft

Ian Soboroff
NIST

1 Overview

The TREC Web Track explores and evaluates Web retrieval technologies. Currently, the Web Track conducts experiments using the new billion-page ClueWeb09 collection[1]. The TREC 2009 track is the successor to the Terabyte Retrieval Track, which ran from 2004 to 2006, and to the older Web Track, which ran from 1999 to 2003. The TREC 2009 Web Track includes both a traditional adhoc retrieval task and a new *diversity task*. The goal of this diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. For example, given the query "windows", a system might return the Windows update page first, followed by the Microsoft home page, and then a news article discussing the release of Windows 7. Mixed in these results might be pages providing product information on doors and windows for homes and businesses.

The track used the new ClueWeb09 dataset as its document collection. The full collection consists of roughly 1 billion web pages, comprising approximately 25TB of uncompressed data (5TB compressed) in multiple languages. The dataset was crawled from the Web during January and February 2009. For groups who were unable to work with this full "Category A" dataset, the track accepted runs over the smaller ClueWeb09 "Category B" dataset, a subset of about 50 million English-language pages.

Topics for the track were created from the logs of a commercial search engine, with the aid of tools developed at Microsoft Research. Given a target query, these tools extracted and analyzed groups of related queries, using co-clicks and other information, to identify clusters of queries that highlight different aspects and interpretations of the target query.

These clusters were employed by NIST for topic development. Each resulting topic is structured as a representative set of subtopics, each related to a different user need. Documents were judged with respect to the subtopics, as well as with respect to the topic as a whole. For each subtopic, NIST assessors made a binary judgment as to whether or not the document satisfies the information need associated with the subtopic.

These topics were used for both the adhoc task and the diversity task. For both tasks, participants executed the original target queries over the ClueWeb09 collection. The tasks differ primarily in their evaluation measures. The adhoc task uses an estimate of mean average precision, based on overall topical relevance [3]. The diversity task uses newer measures, based on the subtopics, which explicitly consider novelty in the result list (intent aware precision [1] and α-nDCG [4]).

[1]Further information on the ClueWeb09 collection is available at `boston.lti.cs.cmu.edu/Data/clueweb09`. We thank Jamie Callan, Mark Hoy, and the Language Technologies Institute at Carnegie Mellon University for creating this valuable resource.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **NOV 2009**	2. REPORT TYPE	3. DATES COVERED **00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE **Overview of the TREC 2009 Web Track**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Microsoft,One Research Way,Redmond,WA,98052-6399**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

	Groups submitting Category A runs	Groups submitting Category B runs	Total groups
adhoc task	13	14	25
diversity task	10	10	18
any task	13	16	26

Table 1: Participation in the TREC 2009 Web track.

A total of 26 groups submitted runs to the track, with many groups participating in both tasks. Table 1 summarizes the participation of these groups. About half the groups worked with the full collection. A few groups submitted runs over both the full (Category A) collection and the Category B collection. This report provides an overview of the track, including topic development, evaluation measures, and results.

2 Topics

NIST created and assessed new 50 topics for the task. Two example topics are provided in Figure 1. Unlike most TREC tasks, NIST initially did not release the full topics to the participants. Instead, the initial release of the topics consisted only of the query fields from the 50 topics. No other information regarding the topics was provided. While the topics were fully defined by NIST in advance of the initial release, the detailed topics were released to the participants only after all runs had been submitted.

For the purposes of the diversity task, each topic is structured into a representative set of subtopics, related to different user needs. As shown in Figure 1, topics were categorized as either “ambiguous” or “faceted”. Ambiguous queries are those that have multiple distinct interpretations. We assume that a user interested in one interpretation would not be interested in the others. On the other hand, facets reflect underspecified queries, with different aspects covered by the subtopics. We assume that a user interested in one aspect may still be interested in others.

In turn, each subtopic was categorized as being either navigational (“nav”) or informational (“inf”). A navigational subtopic usually has only a small number of relevant pages (often one). For these subtopics, we assume the user is seeking a page with a specific URL, such as an organization’s homepage. On the other hand, an informational query may have a large number of relevant pages. For these subtopics, we assume the user is seeking information without regard to its source, provided that the source is reliable.

The subtopics are based on information extracted from the logs of a commercial search engine, and are roughly balanced in terms of popularity. When selecting the subtopics, strange and unusual interpretations and aspects were avoided as much as possible. The set of subtopics is intended to be representative, not exhaustive, with the number of subtopics per topic ranging from three to eight, with a mean of 4.9.

For the diversity task, documents were judged with respect to the subtopics. For each subtopic, NIST assessors made a binary judgment as to whether or not the document satisfies the information need associated with the subtopic. For the adhoc task, relevance is primarily judged on the basis of the description field, but if a document is relevant to any subtopic then it is usually relevant to the overall topic. However, for the diversity task, a document may not be relevant to any subtopic, even if it is relevant to the overall topic.

```
<topic number="19" type="ambiguous">
    <query>the current</query>
    <description>
        I'm looking for the homepage of The Current, a program
        on Minnesota Public Radio.
    </description>
    <subtopic number="1" type="nav">
        Take me to the homepage of The Current, a program on Minnesota
        Public Radio.
    </subtopic>
    <subtopic number="2" type="nav">
        I'm looking for the homepage of The Current newspaper in New Jersey.
    </subtopic>
    <subtopic number="3" type="nav">
        I want to find the homepage of The Current newspaper in Hartford.
    </subtopic>
    <subtopic number="4" type="nav">
        I want to find the homepage of The Current magazine in San Antonio.
    </subtopic>
</topic>


<topic number="21" type="faceted">
    <query>volvo</query>
    <description>
        I'm looking for information on Volvo cars and trucks.
    </description>
    <subtopic number="1" type="nav">
        I'm looking for Volvo's homepage.
    </subtopic>
    <subtopic number="2" type="inf">
        Find reviews of the Volvo XC90 SUV.
    </subtopic>
    <subtopic number="3" type="inf">
        Where can I find Volvo semi trucks for sale (new or used)?
    </subtopic>
    <subtopic number="4" type="inf">
        Find a Volvo dealer.
    </subtopic>
    <subtopic number="5" type="inf">
        Find an online source for Volvo parts.
    </subtopic>
</topic>
```

Figure 1: Examples of full TREC 2009 Web track topics.

```
the current  1   current(100) the(79) radio(33) station(26) 3(26) mpr(26) 89(26)
                 mn(13) thecurrent(13) fm(6) www(6) org(6) minnesota(6)
                 minneapolis(6) com(6)
the current  2   current(99) the(83) newspaper(49) nj(49) of(33) absecon(16)
                 point(16) somers(16)
the current  3   san(100) current(100) antonio(100) the(40) magazine(20)
                 events(20)
the current  4   south(100) jersey(100) newspaper(50) newspapers(50) current(50)
the current  5   galloway(99) current(99) the(49) newspaper(49) of(49) nj(49)
the current  6   current(100) newspaper(66) the(33) articles(33) org(33)
the current  7   current(100) hartford(100) the(50)
the current  8   hartford(100) newspaper(79) current(79) obituaries(19) the(19)
                 articles(19)
```

Figure 2: Output of the clustering algorithm is eight clusters, relating to the query “the current”.

All topics are expressed in English. For the TREC 2009 track, all non-English documents were judged non-relevant, even if the assessor understood the language of the document and the document would be relevant in that language. In order to reduce the influence of Web spam, assessors were instructed that pages with misleading and/or malicious content should be considered “not relevant”.

3 Topic Development Procedure

Topic development this year was guided by real search engine usage in two ways: 1) by queries sampled from search logs, and 2) through subtopic development informed by the output of a clustering algorithm. This section describes this sampling and clustering.

To guarantee having 50 queries that were useable by NIST, we sampled 200 queries from the usage logs of a commercial search engine. The 200 were sampled from a January 2009 query log, which is roughly the same time that ClueWeb documents were crawled. The sampling distribution was to prefer queries of medium popularity, on the assumption that very popular queries are too navigational, therefore less challenging, and very rare queries may contain personally identifiable information, so should not be used in experiments. This gave 200 of what we call *torso* queries, such as “the current” and “volvo”, avoiding very *head* queries such as “ebay” and very *tail* queries such as “nick craswell phone number”. We also implemented an Adult Filter, so the 200 queries were unlikely to be seeking adult content.

A clustering algorithm [5] was run on each of the 200 queries. Using the query as a seed, the algorithm first finds other queries that are likely to occur in the same session, using a large amount of aggregated session data. For example, the seed query “the current” is expanded to “current newspaper” and “the current radio” amongst others. Then the same expansion step is run again, yielding a much wider set of queries, some of which are on-topic (e.g. “current newspaper nj”) and some of which stray too far off-topic (e.g. “prairie home companion”).

The algorithm then induces a graph, where each node is a query from the expanded set, and edges connect two queries if they have clicks on the same URL. This co-click information may mean

that there is no longer a path from the seed query to some off-topic queries (no path from "the current" to "prairie home companion"). Disconnected nodes were removed, to reduce noise.

Finally an agglomerative clustering algorithm is run on the graph. The resulting clusters were provided to NIST for use in topic development, in the form of unigram language models. The clusters may contain omissions, and may contain duplicate clusters, but can be used to ensure that subtopics have good coverage of interpretations that are "nearby in the space of user clicks and reformulations. The eight clusters in Figure 2 were used to develop the subtopics for topic 19, appearing in Figure 1.

4 Adhoc Task

An adhoc task in TREC investigates the performance of systems that search a static set of documents using previously-unseen topics. The goal of an adhoc task is to return a ranking of the documents in the collection in order of decreasing probability of relevance, where the probability of relevance of a document is considered independently of other documents that appear before it in the result list. For each topic, participants submitted a ranking of the top 1,000 documents.

The process of executing the queries over the documents and generating the experimental runs was required to be entirely automatic, with no human intervention at any stage, including modifications to retrieval systems motivated by an inspection of the queries. Group were instructed not to materially modify their retrieval system between the time they downloaded the queries and the time they submitted their runs.

Each group could submit up to three runs for the adhoc task. All runs were judged by NIST assessors. Each document was judged on a four-point scale as being "highly relevant", "relevant", "not relevant", or "not relevant, but reasonable". These relevance judgments were made with respect to the description field of the topic associated with the query. The "not relevant, but reasonable" level means that the document is not relevant with respect to the description field, but may be relevant to some other reasonable interpretation of the query.

Results for the adhoc task are shown in Figures 2 and 3. As with previous TREC adhoc tasks, the primary evaluation measure is mean average precision (MAP). For the results reported in the figure, MAP is estimated by the Minimal Test Collection (MTC) method [3], which is also used by the TREC Million Query Track [2].

5 Diversity Task

The diversity task is similar to the adhoc retrieval task, but differs in its evaluation measures and in its judging process, as described above. The goal of the diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. For this task, the probability of relevance of a document is assumed to be conditioned on the documents that appear before it in the result list.

In all other respects, the diversity task is identical to the adhoc task: The same 50 topics were used. Query processing was required to be entirely automatic. Groups could submit up to three runs, all of which were judged.

Group Id	Run Id	Expected MAP	Expected Precision @5	@10	@20
Waterloo	watwp	0.043362	0.331699	0.347432	0.366025
UWaterlooMDS	WatSdmrm3we	0.034555	0.155000	0.162222	0.177072
unimelb	muadibm5	0.033712	0.255000	0.286667	0.297297
MSRAsia	MSRANORM	0.033240	0.430000	0.380000	0.377338
UAms	uvamrftop	0.033225	0.445000	0.415556	0.382783
msrc	MS2	0.029981	0.425000	0.402222	0.390479

Table 2: Top 6 adhoc task results (Category A). Estimates are calculated using the MTC method [3]. Runs are ranked by expected MAP. Only the best run from each group is included in the ranking.

Group Id	Run Id	Expected MAP	Expected Precision @5	@10	@20
UMD	UMHOOsd	0.047620	0.345772	0.399920	0.409814
UDel	udelIndDRSP	0.047082	0.277171	0.356119	0.389080
uogTr	uogTrdphCEwP	0.045983	0.541884	0.528193	0.522276
NEU	NeuLMWeb600	0.044242	0.395000	0.400567	0.406506
ICTNET	ICTNETADRun3	0.043297	0.442089	0.443574	0.442391
EceUdel	UDWAxBL	0.042454	0.334019	0.331361	0.337146

Table 3: Top 6 adhoc task results (Category B). Estimates are calculated using the MTC method [3]. Runs are ranked by expected MAP. Only the best run from each group is included in the ranking.

5.1 Evaluation Measures for the Diversity Task

The diversity task used two evaluation measures: 1) α-nDCG as defined by Clarke et al. [4], and 2) an "intent aware" version of precision, based on the work of Agrawal et al. [1]. The computation of α-nDCG exactly follows the procedure described in Clarke et al., with $\alpha = 0.5$. We refer the reader to that paper for further information.

For a given topic, the intent-aware measures of Agrawal et al. [1] treat each subtopic as a distinct interpretation of the associated query. Standard evaluation measures are computed separately with respect to each interpretation. A weighted average is then computed across the various interpretations to give intent-aware versions of the standard measures. Agrawal et al. allow probabilities to be attached to each interpretation. For the TREC 2009 Web track, we assume equal probabilities.

More specifically, we compute intent-aware precision at retrieval depth k using the following procedure. Assume there are M topics. Let N_t, $1 \leq t \leq M$ be the number of subtopics associated with topic number t. Let $j_t(i,j) = 1$ if the document returned for topic t at depth j is judged relevant to subtopic i of topic t; otherwise, let $j_t(i,j) = 0$. We then define intent-aware precision at retrieval depth k as:

$$\text{precision-IA@}k \;=\; \frac{1}{M}\sum_{t=1}^{M}\frac{1}{N_t}\sum_{i=1}^{N_t}\frac{1}{k}\sum_{j=1}^{k} j_t(i,j) \tag{1}$$

		α-nDCG			precision-IA		
Group Id	**Run Id**	**@5**	**@10**	**@20**	**@5**	**@10**	**@20**
MSRAsia	MSRAACSF	0.281	0.316	0.365	0.127	0.112	0.108
msrc	MSDiv3	0.268	0.309	0.346	0.127	0.117	0.105
unimelb	mudvimp	0.220	0.241	0.268	0.091	0.073	0.061
THUIR	THUIR09FuClu	0.206	0.234	0.271	0.105	0.094	0.086
uogTr	uogTrDYScdA	0.159	0.191	0.233	0.074	0.077	0.089
utwente	twCSodpRBB	0.144	0.164	0.193	0.067	0.062	0.061

Table 4: Top 6 diversity task results (Category A). Runs are ranked by α-nDCG@10. Only the best run from each group is included in the ranking.

		α-nDCG			precision-IA		
Group Id	**Run Id**	**@5**	**@10**	**@20**	**@5**	**@10**	**@20**
Waterloo	uwgym	0.335	0.369	0.400	0.162	0.144	0.122
uogTr	uogTrDYCcsB	0.253	0.282	0.308	0.142	0.132	0.127
ICTNET	ICTNETDivR3	0.251	0.272	0.301	0.104	0.095	0.092
Amsterdam	UamsDancTFb1	0.232	0.250	0.281	0.086	0.079	0.071
CSIUCD	UCDSIFTdiv	0.212	0.249	0.278	0.112	0.121	0.115
LU_WUME	wume1	0.220	0.247	0.279	0.121	0.113	0.108
NEU	NeuDiv1	0.215	0.243	0.278	0.126	0.131	0.134

Table 5: Top 6 diversity task results (Category B) with uwgym included for comparison (see text). Runs are ranked by α-nDCG@10. Only the best run from each group is included in the ranking.

5.2 Diversity Task Results

Figures 4 and 5 present the top results for the diversity task. Both α-nDCG and precision-IA are reported at retrieval depths 5, 10, and 20. Note that the run with id "uwgym" should not be viewed as an official submission by a track participant (although it does satisfy the rules for a valid submission). The run was generated by one of the track coordinators (Clarke, with the aid of graduate student Hani Khoshdel-Nikkhoo) to act as a baseline run for the track. To generate the run, the queries were submitted to one of the major commercial search engines. The results were then filtered against the URLs appearing in the Category B collection.

Efforts to generate diversity met with mixed success. Overall, 18.1% of subtopics have no relevant documents. Of the 50 topics, only half have at least one relevant document for every subtopic. The evaluation script for the diversity task ignores subtopics for which there are no judged relevant documents.

Diversity and relevance are closely related. Runs that return many relevant documents tend to have higher subtopic recall. Figure 3 illustrates this relationship. The figure shows a scatter plot of subtopic recall at depth 20 vs. precision at depth 20 for the Web Tack runs. For this plot, we compute precision at 20 by ignoring distinctions between subtopics. Relevance judgments for subtopics are combined by treating a document as relevant to a topic if it is relevant to any of its subtopics. We compute subtopic recall as the number of subtopics with relevant documents in the top 20 divided by the total number of subtopics with relevant documents in the collection.

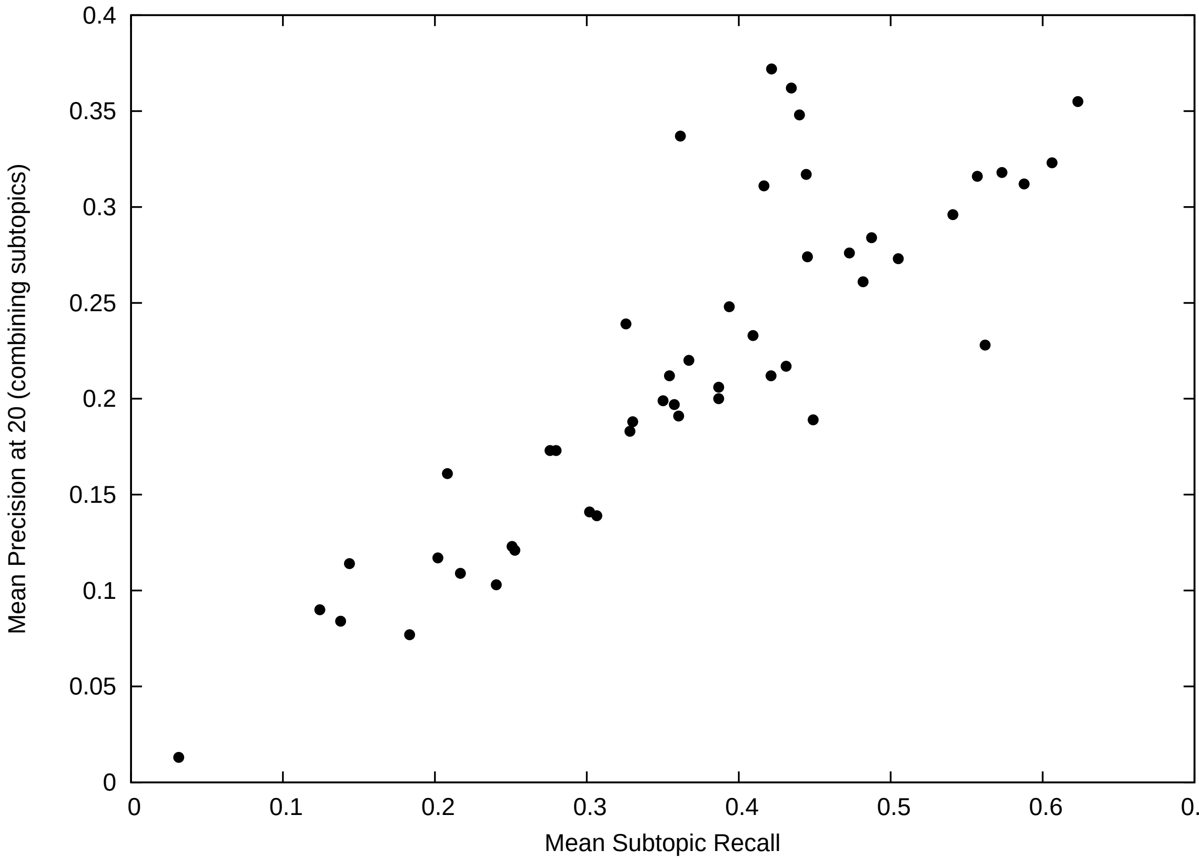

Figure 3: Scatter plot of subtopic recall at depth 20 vs. precision at depth 20 for diversity runs. For this plot only, precision at 20 is computed by combining subtopics, where we treat a document as relevant to a topic if it is relevant to any of its subtopics.

6 Future Directions

The Web Track will continue for TREC 2010. Our current plans include both an adhoc task and a diversity task. In addition, we intend to introduce a Web spam task. As an aid to participating groups, we plan to distribute a preliminary spam ranking of the full Category A collection in early 2010.

Based on discussions during the conference, we intend to make a number of changes to the pooling and judging process. Several groups requested that the top documents from the diversity runs be judged using the adhoc criteria, which would aid them in making comparisons between methods. Other groups requested complete judgments of the adhoc runs to some fixed depth. We hope to accommodate these requests for TREC 2010, resources permitting.

As illustrated by Tables 4 and 5 the top Category A and Category B runs achieve similar performance. This outcome contradicts our expectation that a good Web search engine would to do better on Category A (which includes Category B as a subset) partly because it should benefit from the availability of a larger link graph and more anchor text. It is possible that, due to factors such as crawl order, Category B contains higher quality pages and less spam. The Wikipedia is

included in the Category B collection and may exert some influence on the results. We expect the differences between the categories to be explored by participating groups over the next year. At the present time, we plan to include both Category A and B runs in TREC 2010.

References

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, Barcelona, Spain, 2009.

[2] James Allan, Javed A. Aslam, Ben Carterette, Virgil Pavlu, and Evangelos Kanoulas. Million Query Track 2008 overview. In *17th Text REtrieval Conference (TREC 2008) Proceedings*, Gaithersburg, Maryland, 2008.

[3] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 268–275, Seattle, Washington, 2006.

[4] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkann, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–666, Singapore, 2008.

[5] Filip Radlinski, Martin Szummer, and Nick Craswell. Inferring query intent from reformulations and clicks. In *19th International World Wide Web Conference*, Raleigh, North Carolina, April 2010.